

**National Assessment of Educational Progress (NAEP)
Technology-Based Assessment (TBA) Project
Math Online (MOL) Pretest Report
April 2, 2001**

The Technology-Based Assessment (TBA) Project is intended to facilitate the incorporation of new technology in the NAEP program. The time is right for exploring the use of technology, especially the use of the computer, as a tool to enhance the quality and efficiency of educational assessments. Virtually all of the nation's schools are equipped with computers, and students are using them with increasing frequency and fluency. To help understand how the use of computer technology will impact NAEP assessments, the project attempts to answer the following key questions:

1. *What are the measurement implications of using TBA in NAEP?*
2. *What are the equity implications of using TBA?*
3. *What are the efficiency and cost implications of using TBA?*
4. *What are the operational implications of TBA?*
5. *How can we best incorporate new technology into NAEP in both the short and long term?*

To answer these questions, NAEP is currently undertaking three empirical studies: Mathematics Online (MOL), Writing Online (WOL), and Problem Solving in Technology-Rich Environments (TRE). For each of these studies, we are using the same basic development cycle consisting of a small-scale pilot test (about 50 students), a pretest (between 100 to 500 students), and a full-fledged special study (a sample of 1,000 students). This three-tiered development cycle allows us to create, try out, and refine our online measures iteratively, first piloting with a very small sample of students and schools, then resolving scale-up issues with a somewhat larger pretest sample, and finally collecting data with a field-test sample large enough to support statistical inferences. It is this larger sample that will enable us to answer some of the questions above in a rigorous way.

This report concerns the pretest phase of MOL, conducted in April and May 2000. The MOL pretest was designed to explore how our software and logistical procedures scaled up from the earlier pilot samples and to give an initial sense of how students in fourth and eighth grades were able to respond to mathematics questions delivered on a computer. Keep in mind that the MOL pretest samples at fourth grade ($N = 95$) and eighth grade ($N = 183$) are not adequate to answer questions about aspects such as item performance and equity that require larger and more representative samples. Data from the subsequent field trial will be used to answer these important questions. However, the NAEP Program has learned a number of valuable lessons from this phase of the study. Because these lessons may be of use to others interested in computer-based testing, a series of interim reports will be issued. This paper represents the first of such interim reports.

Goals of MOL

The primary goals of the MOL pretest phase were to gain insights into the operational and logistical mechanics of computer-delivered assessments and to evaluate the ability of fourth- and eighth-grade students to deal with assessments delivered on computer. In particular, there was concern about the degree to which the computer skills of fourth- and eighth-grade students might impact performance. The concern was greatest for fourth-grade questions that required students to type explanations or draw figures as components of constructed-response questions.

Sources of Items

All questions used in MOL at the fourth and eighth grades were converted into a computer-deliverable format from existing fourth- and eighth-grade paper-and-pencil NAEP instruments. Multiple-choice and constructed-response items were included at both grade levels. In the paper-and-pencil instruments, students filled in ovals for multiple-choice items and wrote their answers and explanations for constructed-response items. For questions administered online, students were required to “click” on the correct answer choice for multiple-choice items. For constructed-response items, they were required to enter their answers in different ways, depending upon the item type. Response modes included typing a number or an explanation, drawing a figure with the mouse, or building a figure by clicking on line segments.

Description of MOL Instruments

The fourth-grade MOL consisted of the following components.

- A tutorial was designed to train students on how to use the computer to complete the assessment. The tutorial had an embedded typing quiz and tasks to provide a measure of keyboarding skill. The tutorial also provided practice in clicking on choices, clicking to shade or darken regions, moving back and forth between screens, correcting errors, and typing answers and explanations. The tutorial was designed to be completed in approximately 10 minutes.
- Cognitive questions were presented in three sections, the last of which was preceded by a tutorial to teach students how to use an online 4-function calculator needed for some of the items in that section. There were a total of 31 questions: 22 multiple-choice, 7 short-constructed response, and 2 extended-constructed response that required an explanation.
- Background questions presented in one section were designed to gather information about students’ demographic background and computer experience. There were 32 questions in this section.

The eighth-grade MOL consisted of the following components.

- A tutorial designed to train students on how to use the computer to complete the assessment. The eighth-grade tutorial covered the same material as the fourth-grade tutorial, with one additional exercise added to teach students how to use a response template dealing with whole numbers, decimals, fractions, and mixed numbers.
- Cognitive questions presented in three sections, the last of which was preceded by a tutorial to teach students how to use an online scientific calculator. There were a total of 26 questions: 16 multiple-choice, 6 short-constructed response, and 4 extended-constructed response.
- Background questions presented in one section were designed to gather information about students' demographic background and computer experience. There were 33 questions in this section, most of which were the same as those used in fourth grade.

School Recruiting and Technical Certification of School Computers

The goal was to administer MOL pretests at fourth grade in 10 schools with 10 students from each school and at eighth grade in 20 schools with 10 students from each. In order to simplify administration and technical support, schools recruited came from a pool of 298 schools, located in Maryland, New Jersey, New York, North Carolina, Pennsylvania, Virginia, or Washington DC, most of which had participated in other NAEP assessments from January to March, 2000. A total of 36 schools with 95 students at fourth grade and 183 students at eighth grade participated.

In order to participate in the study, schools needed to meet NAEP-specified computer hardware and Internet connection specifications. For this study, the specifications included:

- Computers: Pentium class machine with 166 megahertz microprocessor or better; 32 MB of RAM; CD-ROM drive
- Monitors: 17-inch monitor with 1024 x 768 screen resolution and 65,536(16-bit) colors minimum
- Software: Windows 95, Windows 98, or Windows NT, Internet Explorer 5.0
- Internet connection: 128k bit or better dedicated line

These specifications were created to allow use of tests with advanced features, which could be delivered over the web for MOL as well as for other TBA projects. These advanced features included on-the-fly (OTF) item delivery (for MOL), innovative items involving drawing responses (for MOL), full-featured word processing capability (for WOL), self-voicing of directions and items (for WOL), and enhanced color and graphics (for TRE). These features were selected to insure that the TBA studies explore capabilities that will be commonly available by the time NAEP is ready for operational use of computers. However, the hardware and software required to support these features are not yet widely available in schools, and this led to operational challenges.

Westat staff contacted schools in a two-step process. Initially, Westat contacted school staff by telephone to determine whether the computers and Internet connection at the school met the minimum standards required by NAEP. Schools that indicated they met the standards for the project were subsequently visited by Westat or ETS test administrators and technical support staff, who installed software for MOL and adjusted the settings on the school computers in preparation for MOL.

Key findings related to recruiting and technical certification included the following:

- Most schools contacted were interested and enthusiastic about participating.
- Of the schools expressing interest, a large percentage of schools were disqualified from participation because they had only Macintosh computers, had inadequate or no Internet connection available to students, or had PCs that did not meet the technical specifications required for the study.
- Many schools lacked knowledgeable in-house computer technical staff or any other staff who could determine to what degree the school met the technical specifications called for in the study. At some schools, staff provided inaccurate information about the school's computers or Internet connection, causing problems when Westat staff attempted to complete software installation or tried to complete test administrations.
- Installation required more ETS and Westat staff time than could be expended in a field test or operational NAEP assessment.

These lessons led to several adjustments in program plans. First and foremost, ETS and Westat worked out a stand-alone delivery system that could be delivered on laptop computers so more schools could participate. Second, we realized the fiscal and operational importance of obtaining accurate information about school equipment without having to make expensive visits to schools. ETS has now developed a way to determine, without having to visit a school, if the particular computers in a school meet the delivery requirements of MOL and related assessments. Remote certification was not in place for this phase of the study, but will be used in the spring 2001 MOL field test.

Test Delivery and Administration

Three test delivery models were tried: Internet delivery using school computers connected to the Internet, laptop delivery using Westat-supplied stand-alone computers not connected to the Internet, and school computers connected to a Westat-supplied laptop server. Westat administrators found the stand-alone laptop delivery mechanism to be relatively simple to use. This delivery mode and the direct Internet delivery mode will both be used in the spring 2001 MOL field test. The in-school server approach will not be pursued because it required advanced technology skills not normally expected of Westat test supervisors.

Key findings related to test delivery and administration included the following:

- Students and administrators reacted to the test favorably at both the fourth and eighth grades. Test administrators felt that the students enjoyed completing the tutorial and taking the test. The tutorials seemed to accomplish their intended purpose. While these observations are limited and anecdotal, they do suggest that the larger study will be well received.
- Westat and ETS realized the need for a test delivery system that could be easily managed without a need for the assistance of on-site computer technical support staff. ETS is working with Westat on simplifying the technical requirements for NAEP TBA computers and, as mentioned above, has developed a “remote certification” system that participating school personnel can use independently to determine whether their school should complete MOL using school computers or Westat-provided laptops.
- Providing scrap paper and pencils (rather than providing an electronic form of scrap paper) for students to use during the assessment seemed to work well for students taking the assessment.
- Westat test administrators felt comfortable administering the test on laptop computers because little technical skill is needed.
- The most frequently encountered technical problems were with the schools’ Internet connections, with the ETS server, or with the schools’ PCs. There were also occasional problems with schools’ Internet “firewall” security systems, which prevented the MOL software from running.
- At eighth grade, test administrators favored giving the test to students individually as they arrived at the test center, rather than conducting group administrations.

Item Development and Formatting

The items used for MOL at the fourth and eighth grades came from previously developed NAEP sources. The fourth grade MOL items were taken from the “Market Basket” assessment administered in conjunction with the NAEP 2000 mathematics assessment. Eighth-grade items were selected from the existing pool of eighth-grade NAEP mathematics items so that the MOL instrument was broadly representative of the NAEP math framework content and item type specifications.

The process of translating items for computer delivery included: (1) revising paper-and-pencil directions to be appropriate for online presentation; (2) designing the screen layout, appearance, and usability of item formats so that students could complete the assessment with minimal instruction; and (3) finding ways of authentically assessing math framework content while using online delivery formats.

The formats of some of the paper-based items were modified slightly to adapt to online administration while retaining essential elements of content. For example, for one item at eighth grade, the dimensions of a diagram were reduced to allow presentation of the entire graphic on the screen because the original dimensions would have forced students to scroll to view all parts of the stimulus.

Some items were dropped and/or replaced because they were judged too problematic to put online. For example, one dropped item called for students to draw a triangle to scale using a ruler. Testing this kind of skill on the computer seemed problematic in that using a physical ruler is not the same as using a virtual ruler. Also, a higher level of computer skill would be required on the part of the student to manipulate the virtual ruler.

At fourth grade, three of the 31 items originally selected were dropped from MOL. One three-part pencil-and-paper item was reformatted into three separate items for MOL. (Interestingly, MOL performance on these three items was uniformly higher than for the regular NAEP sample, but the small sample size does not permit us to draw a firm conclusion about the meaning of that difference.) At the eighth and sixth grades, 6 of 28 items were dropped and replaced with 4 other items, resulting in a total of 26 items.

Based on the limited sample of fourth- and eighth-grade mathematics blocks from which MOL items were selected, items assessing three of the mathematics framework dimensions were implemented relatively easily: (1) *Number Sense, Properties, and Operations*, (2) *Data Analysis, Statistics, and Probability*, and (3) *Algebra and Functions*. Two framework dimensions proved more difficult to implement: (1) *Measurement* and (2) *Geometry and Spatial Sense*.

Key findings related to item development and formatting included:

- Certain mathematics framework items in given item formats may present significant validity or logistic problems for online presentation and may be best delivered in paper-and-pencil blocks, rather than online. For example, some items that require students to make a scale drawing or to work with geometric shapes might be best administered on paper. On the other hand, some mathematics skills may be more easily and accurately assessed using computers rather than paper-based tests.
- One of the most complicated tasks called for in math items is asking students to show their work or to explain their answers. Newly developed items need to consider the appropriate use of online presentation for this framework requirement. For example, in mathematics, students may choose to explain their answer by making an annotated drawing. Presenting such response formats with current technology can be problematic, although advances in graphics capture are occurring that should make this response format viable in the future.
- With a few exceptions, students at fourth grade were able to use the printed handouts and rulers associated with a small number of MOL items. For these items, students viewed the question on the computer, measured a printed stimulus (e.g., a geometric figure) using their ruler, and then typed their answers on the computer. (No such items were tested at eighth grade).
- The online calculators and calculator tutorials worked well for students at both the fourth and eighth grades.

Of the four basic item format types included in MOL, initial impressions of the efficacy of each are as follows:

- *Single selection multiple-choice*
Worked very well for students at the fourth and eighth grades.
- *Very short answer (type in a number, a word, or a sentence)*
Worked well for students at the fourth and eighth grades.
- *Click on several graphics to create an answer*
Worked well for students at the fourth and eighth grades.
- *Click or type an answer, and then type an extended explanation.*
Worked well for students at eighth grade. Some evidence at fourth grade (two items) suggests that students struggled with typing fairly long explanations.

Scoring Students' Responses

Students' multiple-choice responses were scored by computer. Students' constructed-response responses were scored by NAEP test development staff. In addition, responses for one constructed-response item were scored automatically. This item allowed the student to draw a geometric figure freehand with the mouse.

Analysis of the students' responses indicated the following:

- A very small number of apparent student keyboarding errors (for example, failure to move the cursor to a different text box, unintelligible text typed in answer boxes) were observed at the fourth and eighth grades.
- At eighth grade, students who chose to use the mouse to draw a geometric figure freehand for one item were reasonably successful.
- In the short term, it is likely that students' extended-constructed responses will need to be evaluated by human scorers.
- Based on an initial examination, the potential for automatically scoring certain types of students' short-constructed responses (e.g., those that can be answered with a single number, word, or simple geometric construction) is promising. ETS staff is currently actively exploring the development of methods to score phrases and sentences, so that, in the longer term, the automatic evaluation of these responses should be possible.

Because the spring 2000 MOL pretest was limited to a small sample size and was not nationally representative, no comparisons between student performance on MOL and paper-based tests can be drawn. (An examination of subgroups' performance and comparisons across delivery mode will be made using the data collected in the MOL field test to be administered in spring 2001.)

On-the-Fly Testing

At eighth grade, an additional goal of the study was to investigate the technical feasibility of generating alternate versions of multiple-choice and constructed-response items using an “on-the-fly” (OTF) technology. Test items generated using OTF technology were based on models derived from original, or “parent” versions of items. The OTF models were created by varying elements in the parent items and by setting constraints on how these variable elements related to one another. Parallel versions of parent items were then automatically generated in real time and presented to students. This was done for 11 items (8 multiple-choice and 3 constructed-response) from the eighth-grade test. The test administration procedures and the ordering of items in the eighth-grade OTF test were identical to that of the “regular” eighth-grade MOL test. Students were assigned randomly to a “regular” test form or to an OTF test form.

Findings with respect to this aspect of the study were:

- Items for which it proved to be too difficult to produce on-the-fly (OTF) versions included those in which the multiple-choice options were geometric figures or those that required the modification of geometric figures. (We have recently implemented a mechanism for automatically rendering figures and figure labels on the fly. This mechanism is now being pilot tested in another mathematics project. Although it will not be available in time for the MOL field test, this mechanism should allow the on-the-fly generation of such items in future NAEP activities.)
- There were a few software problems initially delivering the MOL OTF test at eighth grade, so fewer OTF tests than anticipated were administered. Later in the pretest, there were no reports of problems from either students or test administrators.
- There is a trade-off between content coverage, calibration, and scoring efficiency in the use of OTF items. The principle behind OTF is that item variants should allow for increased content coverage for a framework element. Efficiencies in calibration occur because item models, not individual item variants, are calibrated. The trade-off occurs because, depending on the aspects of the items that are varied, scoring the variants from an OTF constructed-response item family may prove more time-consuming than scoring the original parent constructed-response item. This increase occurs because the diversity of responses encountered in alternate versions of the question will require scorers to review more training samples and to work more of the questions through in order to score them. Such a trade-off obviously does not occur for automatically scorable on-the-fly items.

Implications for the Future

As more staff members gain experience working on TBA projects, and when MOL is administered to a larger sample of students in spring 2001, more conclusive results can be

reached about the efficacy of MOL for NAEP. However, based on the results to date, several general implications for the future of NAEP TBA projects are suggested.

Scheduling and planning

- Some development procedures that are currently easy to do for paper-and-pencil tests (for example, reviewing and revising items) need more time or tend to be more complicated for items delivered online.
- For operational NAEP, there are very short time windows between all of the steps in any assessment project (question tryouts, committee meetings, clearance, post-clearance revisions, training, administration, scoring, and reporting). Schedules for online items need to be developed taking into account the complexity that online testing may add to all parts of an assessment.

Item development, review, and revision

- Development of future test items will depend on close collaboration between software developers, test specialists, and committee members. It is extremely important that technical and operational staff be involved in the very earliest stages of test design and item development in order to ensure that items can be implemented effectively. Those who write and review items clearly need to be alerted to which types of items seem to work well in online delivery mode and which do not.
- Online test directions (animated tutorials and demonstrations) will need to be developed at the same time as the items.
- Future test development efforts for eighth-grade students could include online rulers, protractors, and other tools. The desire to implement such items online should be balanced against the cost of development, tutorial time needed to teach students how to use the tools, and faithfulness to the math framework dimension being assessed.
- As new online item types are developed, usability testing of online items will need to be included as a regular part of the question tryout process currently employed in NAEP for cognitive paper-and-pencil items. Scheduling this process will be much more difficult than with paper-and-pencil items.

Test administration and delivery

- In order to ensure nationally representative data for assessments conducted in the near future, NAEP computer-based tests may need to be delivered via laptop, in cases where school technology infrastructure cannot support the delivery of the test via the Internet.

Concluding Remarks

This report covers the pretest phase of the National Assessment of Educational Progress (NAEP) Mathematics Online (MOL) study. The MOL study is part of a series of investigations intended to facilitate the incorporation of new technology in NAEP.

The MOL pretest was conducted to assess how effectively our software and logistical procedures for delivering technology-based tests (TBA) in fourth- and eighth-grade mathematics scaled up. Among the most important results was that fourth- and eighth-grade students appeared to enjoy interacting with the computer-based assessment software and seemed generally able to use it to provide answers to questions presented online. With respect to logistics, we found it difficult to get reliable information about hardware and software capabilities at individual schools, suggesting that a remote querying process that determined whether school technology infrastructure was suitable for Internet test delivery might be useful. In addition, we discovered that the existing school technology infrastructure would not allow us to depend on Internet delivery alone. Rather, using school machines connected to the Internet and bringing NAEP laptops into those schools that cannot employ Internet delivery should permit us to collect data from a representative sample of schools in our upcoming field test. That field test, which will provide the data needed to answer fundamental questions about the value of technology-based assessment for NAEP, is being conducted this spring.